Section 0: Brief overview- Research proposal and scope of Bio-statistics in quantitative research

Learning objectives This morning?

- Everyone should be able to name and describe what a research proposal should contain
- Be able to identify and appreciate the importance of Statistics in any research endeavor

What is a research ?

A systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions

Characteristics of research

The characteristics of research include all of the following :

- It is the **systematic collection**, **analysis** and **interpretation** of **data** (*need Biostatistics knowledge*)
- It demands a clear statement of the problem
- It requires a **plan** and must be **organized**
- It builds on **existing data**, using both positive and negative findings

What is a research proposal

Proposition to do research

What is it we want to study

What is the question we want answered

How are we going to have this question investigated and answered

THINK OF THE READER

Where do you start??

You start with a RESEARCH general TOPIC in mind

Be careful: Just having a topic is not enough

- For example collecting empirical data to answer a question on a non specific and broad topic is an up hill task

Very often lack of adequate knowledge on what is known is a big problem to students

READ READ READ About your General topic

Literature …

At least someone has written some article or book etc. AROUND your topic

If you do not want to read, please say bye bye to research

It is key in getting a focused informed topic .

Eases the way to a good research question

Research question-Research definition

Sometimes it involves formulating the problem statement which begs for research questions

It might be that you start with a CENTRAL Question and then follow with a series of investigating questions

Conceptualise a model for testing

It could be a research hypothesis

It could take other forms..

-ONLY AFTER LITERATURE REVIEW

-

You must clarify how you would answer your question-Methods

General research design

What type of data we are collecting and how we collect the data

Develop the instruments

From who we are collecting the data(population, sample)

Data collection procedure

Planned data analysis

Ethics

Budget

Development of a research idea

- The order recommended in the development of a research idea
  - **Research topic,**
  - **research problem,**
  - **research purpose,**
  - **research question,**
  - **hypothesis**

The Scientific Method

1. Raising a Question from observation.
2. Suggest Hypothesis.
3. Literature Review.
4. Literature Evaluation.
5. Acquire Data and Data Analysis. (**Experiments**)
6. Data Interpretation.
7. Hypothesis Support.
8. Reporting (**Draft Paper for stakeholders, conference and publication**)

Structure of a Thesis Proposal

1. Title page
2. Abstract
3. Table of contents
4. Introduction
5. Thesis statement
6. Approach/methods
7. (Preliminary) Results
8. Discussion
9. (Work plan including time table)
10. Implications of research
11. List of references
12. Appendix

Courtesy, Prof Mbacham

The IMRAD Format

Papers of original research are written using the IMRAD format (Introduction, Methods, Results, and Discussion).

However, different formats are used for a **review paper**

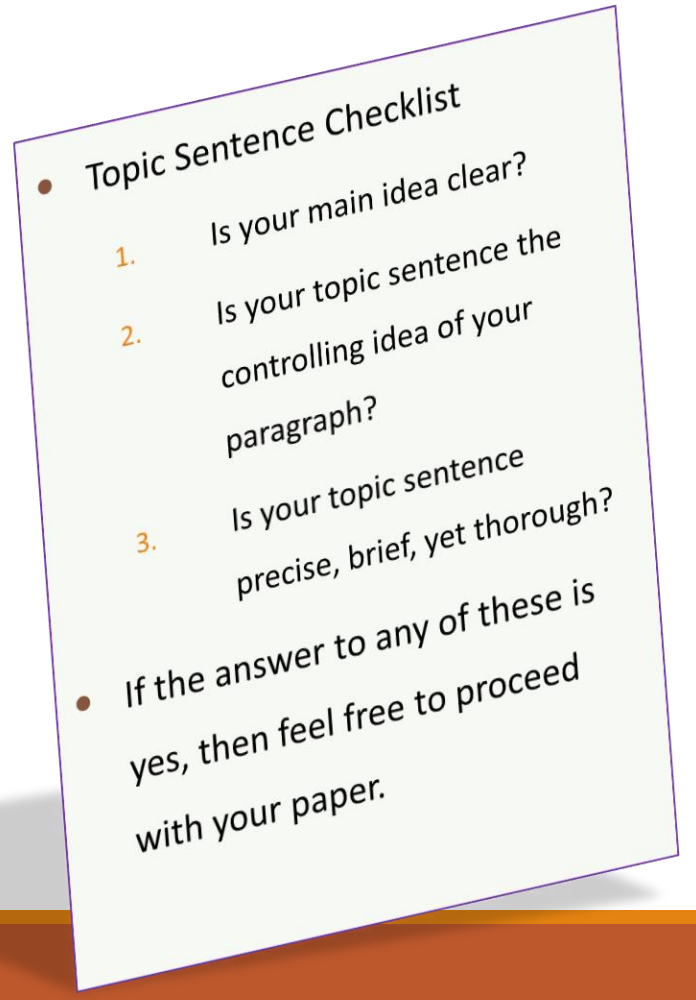
Each section contributes to the overall story by answering one or more questions.

Title

Characteristic of a well-written abstract of a paper:

Concise with about 21 words maximum

Tell the story of the paper



ABSTRACT

Clinical Chemistry 56:4
521–524 (2010)

Clinical Chemistry
Guide to Scientific Writing

The Abstract and the Elevator Talk: A Tale of Two Summaries

Thomas M. Annesley

What is an elevator talk, and what does it have to do to the reader, with the goal of enticing the reader to

Characteristic of a well-written abstract of a paper:

- Stands on its own without need to read the paper
- States the hypothesis, question, or objective of the study

ABSTRACT

Characteristic of a well-written abstract of a paper:

- Does not include references to support the hypothesis, question, or objectives
- Does not contain information absent in the paper
- Does not cite tables or figures

Abstract ex: Any Weakness?

BACKGROUND: Atherosclerotic disease is a major cause of death in the United States. We investigated which analyte, IL-6 or β-selectin, would be a better prognostic marker for atherosclerotic disease.

METHODS: We divided patients into 4 groups. Specimens from each patient were tested for interleukin-6 and β-selectin and matched against the patient's disease group. During the study period, these analytes were measured again to determine whether concentrations changed with disease severity. Mortality was also monitored for each group to investigate any relationship

between IL-6 or β-selectin and the risk of death.

RESULTS: The IL-6 concentrations were different between groups, with the IL-6 concentrations significantly different between groups 1 and 3, and 1 and 4. Although IL-6 and β-selectin concentrations both changed, β-selectin changed by only 10% to 30%. Changes in disease severity were reflected in changes in IL-6. IL-6 values were the same for men and women and did not show any relationship with patient age. Intraindividual variation for IL-6 was much lower than that for β-selectin.

CONCLUSIONS: IL-6 and β-selectin concentrations change with a change in heart disease severity. Intraindividual variation of IL-6 was also much lower than β-selectin, further validating the use of IL-6 over β-selectin. Further work is needed to confirm this observation.

Introduction

Should provide answers to:

- **What problem, question, or hypothesis is being studied?**
- **Why would it be of interest to the reader?**

Clinical Chemistry 56:5
708–713 (2010)

Clinical Chemistry
Guide to Scientific Writing

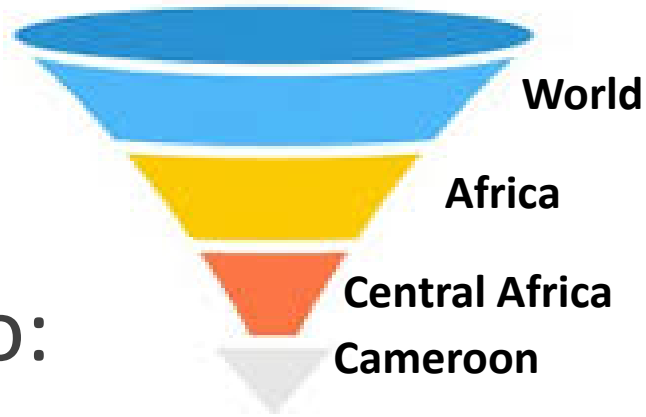
"It was a cold and rainy night": Set the Scene with a Good Introduction

Thomas M. Annesley

In theatrical productions, there is a process called *set-*

pectancy, medical costs, etc., of cancer in general. Get

Introduction



Should provide answers to:

- **What problem, question, or hypothesis is being studied?**
- **Why would it be of interest to the reader?**

The information in the Introduction flows from broad to narrow. The first paragraph provides general background material on the topic, and the last paragraph focuses on the specific question(s) being asked in the study.

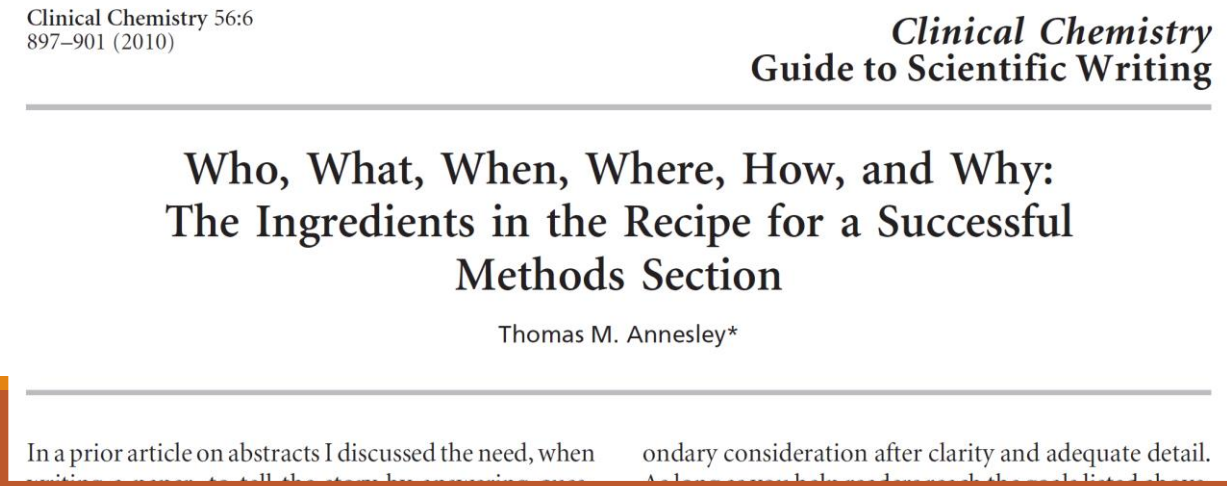
Statement of the problem

One sentence that explains the rational of the work

Methods

Should provide answers to:

- **How did you perform the study?**
- **How did you test the hypothesis, or answer the question?**



Clinical Chemistry 56:6
897–901 (2010)

Clinical Chemistry
Guide to Scientific Writing

Who, What, When, Where, How, and Why:
The Ingredients in the Recipe for a Successful Methods Section

Thomas M. Annesley*

In a prior article on abstracts I discussed the need, when

ondary consideration after clarity and adequate detail.

Methods

Should provide answers to:

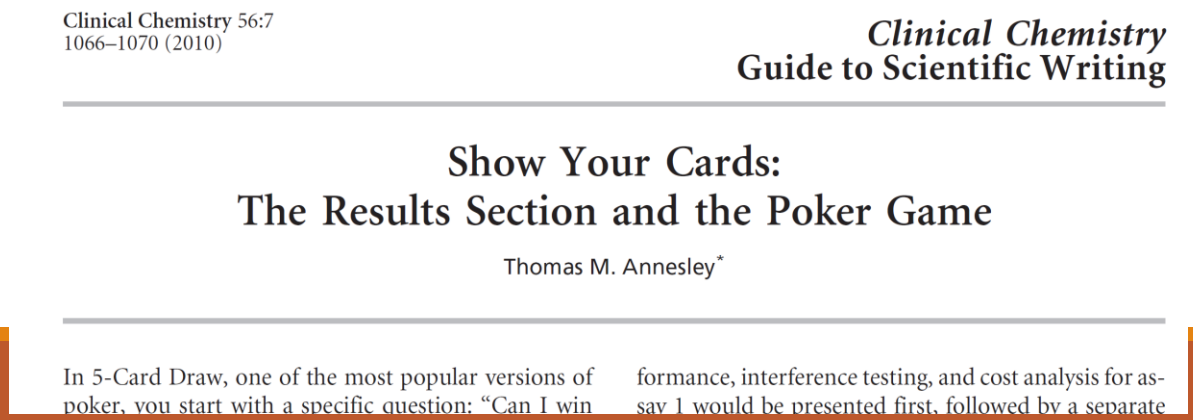
- Describe in details the research participants
- The research sites
- The inclusion criteria
- Non inclusion criteria and
- exclusion criteria

Results

Should provide answers to:

- **What did you find?**

- **Did you solve the problem, prove the hypothesis, or answer the question?**



Clinical Chemistry 56:7
1066–1070 (2010)

Clinical Chemistry
Guide to Scientific Writing

Show Your Cards:
The Results Section and the Poker Game

Thomas M. Annesley*

In 5-Card Draw, one of the most popular versions of poker, you start with a specific question: "Can I win

formance, interference testing, and cost analysis for as-say 1 would be presented first, followed by a separate

Results- Graphs

Choose carefully how to present your results.

Graphs

- Have an immediate visual impact
- are good for showing trends or patterns
- are good for highlighting differences between sets of data.
- do not work well when the exactness or precision of the data is important.

Results- Tables

Choose carefully how to present your results.

Tables

- are better when the individual or summarized values are more important than trends.
- can be used for presenting both quantitative and qualitative data.
- can contain words, symbols, numbers, or a combination of all three.
- allow side-by-side comparison of data

Results- Making effective use of both data and results

Baseline median IL-6 concentrations were 12, 26, 96, and 144 g/L for categories 1 to 4, respectively, and were not found related to age or sex. Median β-selectin concentrations increased 30% across the 4 categories. Increased disease severity and mortality were associated with higher IL-6 concentrations, but not β-selectin. Intraindividual variation for group 1 was 14% for IL-6 and 36% for β-selectin.

Results- Making effective use of both data and results

Baseline median IL-6 concentrations were 12, 26, 96, and 144 g/L for categories 1 to 4, respectively [DATA], and were not found related to age or sex [RESULT]. Median β-selectin concentrations increased 30% across the 4 categories [RESULT]. Increased disease severity and mortality were associated with higher IL-6 concentrations, but not β -selectin [RESULT]. Intraindividual variation for group 1 was 14% for IL-6 and 36% for β-selectin [DATA].

Discussion

Should provide answers to:

- **What do your results mean?**
- **What value do they add to the scientific literature?**

Clinical Chemistry 56:11
1671–1674 (2010)

Clinical Chemistry
Guide to Scientific Writing

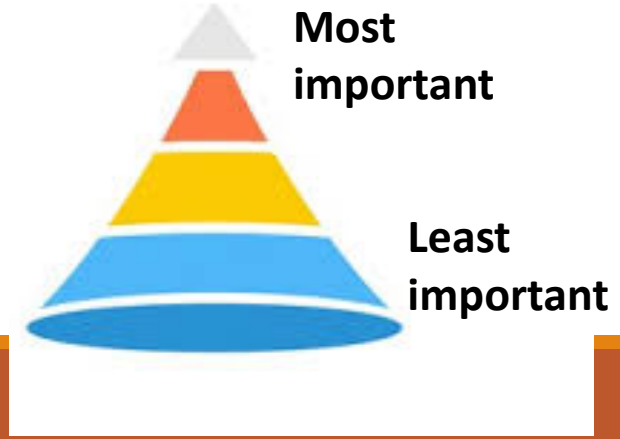
The Discussion Section: Your Closing Argument

Thomas M. Annesley*

Discussion

A good discussion should be:

- **Very specific and focused.**
- **This goal is accomplished by getting right to the point, which is to answer the questions(s) presented in the Introduction?**



Pay attention to

Testing hypothesis use **Inferential statistics which is different from** Descriptive statistics

The feasibility of a research study which should consider:

- a. Cost and time required to conduct the study
- b. Skills required of the researcher
- c. Potential ethical concerns

Main Research methods

Quantitative in nature

- Research measures and explains effects, social events and human behavior with numbers.
- Statistics is key here in analyzing and explaining the findings
- It could be even with the use of Secondary data(Secondary analysis)

Qualitative in nature

- Uses systematic observation and focuses on opinions and what is the hidden meaning of social actions.

Qualitative Research methods

Ethnographic Interview

- Researcher learns about people and their behaviours by talking or conversing with them

Participant observation

- Researcher participates in the activity of the community he/she is learning about while observing the people she is studying.

Case studies

- Intensive observation of a particular person , group or event

Triangularisation: Using many techniques to gather research data.

Content Analysis: Examines and analysises communications

Mixed Research methods

Mixed method simply means in one research research we use both the quantitative and qualitative methods

Qualitative Vs Quantitative Research

QUALITATIVE

Primarily exploratory

Can find out opinions, reasons and motivations

Could give information for quantitative research

Methods –focus group discussion, key informants, participant observations

QUANTITATIVE

Can give rise to quantification of problem or effects or relationships

Uses measurable data to uncover patterns

Methods-longitudinal studies, surveys, cohort, case control studies

For this course!!!-Only Quantitative Research would be assumed

Examples of Research Questions

• What is the efficacy of chloroquine in curing patients with Novel coronavirus COVID-19

•Is artemether Lumefantrine superior to artesunate amodiaquine in the treatment of *P. falciparum* malaria?

•Comparing nutrient content in Foods in Cameroon

A research Project is needed to answer our research question

Steps in a Research Project…

- ✓Planning/Study Design
- ✓Data collection
- ✓Data analysis
- ✓Data analysis presentation
- ✓Interpretation

Scope of Statistics in a Research Project

✓When do you need biostatistics/Statistics??

~~Come and do the magic when I have collected the data!!~~

Biostatistics and or a biostatistician plays a role all through the steps!!!!

Scope of Statistics in a Research Project: Planning/Design of study

✓Setting up main research question: Quantifying single group information and multiple groups

✓ How many participants do you need to answer your research question

✓ How would you select your participants

✓ If you are comparing groups, how do you assign people to a group.

Scope of Statistics in a Research Project: Data analysis

- ✓How best to summarise the data
- ✓ How would you deal with variability in the data
- ✓ How do you use this single study to make a statement about the whole population- Inference

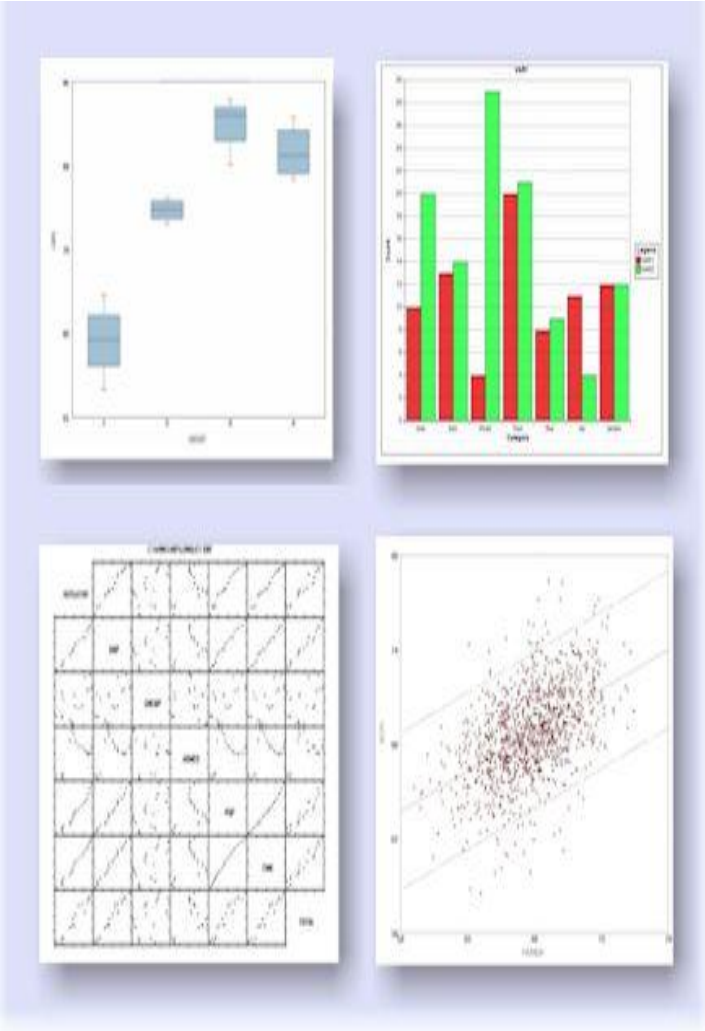
Scope of Statistics in a Research Project: Data analysis Presentation

✓What is the best summary measure to convey the main message from the data .

✓ How do I convey uncertainty estimates in the data

Scope of Statistics in a Research Project: Data analysis Interpretation

✓What does the result mean in terms of the population, patient care and even programs.



What is statistics?

It is a discipline concerned with:

•Designing experiments and other data collection,

•summarizing information to aid understanding,

•drawing conclusions from data,

• and estimating the present or predicting the future.

What is Biostatistics ?

Definition: When the different statistical methods are applied in biological, medical and public health data they constitute the discipline of Biostatistics..

Why Statistics?

✓The world is inundated with data

✓ An increase in Big data sources with technology progression

✓Data is not information until it harnessed

Why Statistics?

✓Headline from a Havard Business Review (Davenport et al , 2012)

“ Data Scientist: The sexiest Job of the 21st Century

Why Statistics?

✓Statistics not only important in scientific world

The Washington post, august 5, 2009 reported..

“ The program conducted last year at 8 high schools found that 13% of about 3000 students tested positive for an STD, mostly gonorrhoea and Chlamydia according to the DC department of Health.”

Why Statistics?

✓Data can be analysed to provide life changing solutions in our world today- Clinical Trials

✓Interpretation of poorly analysed data can cause policy changes that might be harmful to the population.

What is in the mind of a Researcher???

✓ Any Research Project start up with Research Question?

Some questions at the back of the mind

- ✓Why?
- ✓How?
- ✓What?
- ✓When?

What is a Research question?

- A research question is a formal statement of the **goal of a study**

- The research question states clearly what the study will investigate or attempt to prove

- It is a logical statement that progresses from what is known or believed to be true (as determined by the literature review) to that is unknown and requires validation.

Examples of Research Questions

- Is first pregnancy associated with Caesarean Section

- Is artemether Lumefantrine superior to artesunate amodiaquine in the treatment of *P. falciparum* malaria?

- Do patients have a better prognosis if there are multiple rounds by health personnel at their bedsides?

Population and Sample

What is a Population and what is a Sample

Researchers tend to study samples and not the population: WHY???

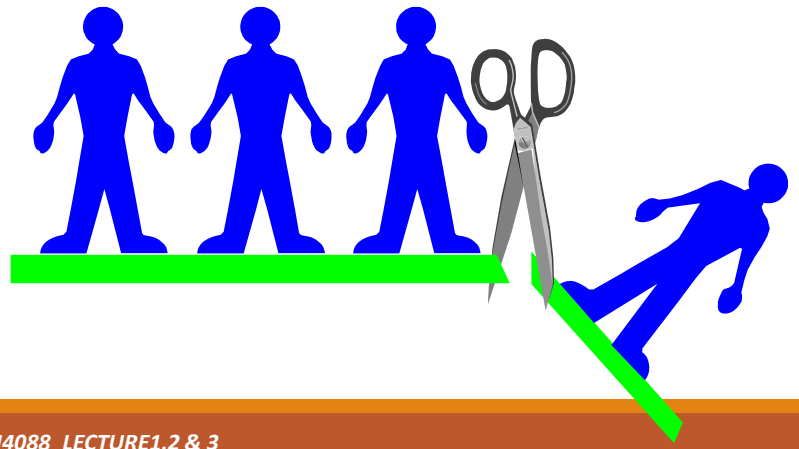
Section1:Sampling

What is sampling?

- If all members of a population were identical, the population is considered to be ***homogenous***.

- That is, the characteristics of any one individual in the population would be the same as the characteristics of any other individual (little or no variation among individuals).

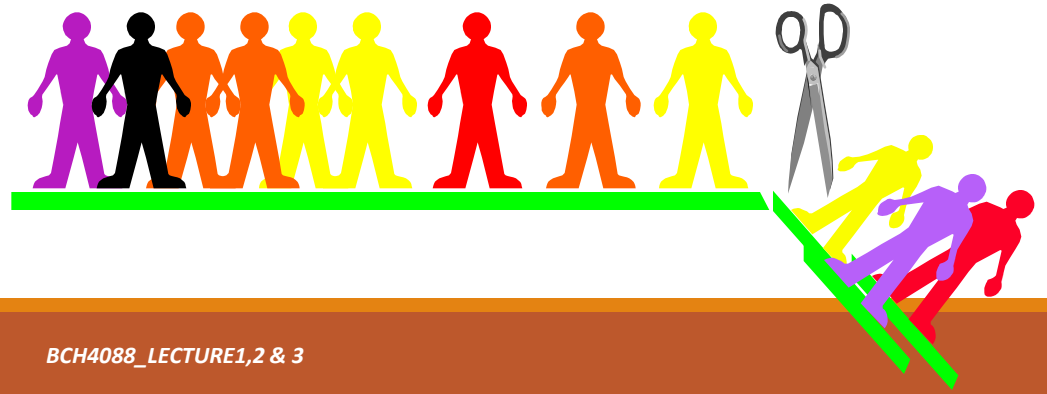
So, if the human population on Earth was homogenous in characteristics, how many people would an individual need to observe in order to understand what humans were like?



What is sampling?

- When individual members of a population are different from each other, the population is considered to be ***heterogeneous*** (having significant variation among individuals).

- In order to describe a heterogeneous population, observations of multiple individuals are needed to account for all possible characteristics that may exist.



Random vs Nonrandom Sampling

Random sampling

- **Every unit of the population has the same probability of being included in the sample.**
- **A chance mechanism is used in the selection process.**
- **Eliminates bias in the selection process**
- **Also known as probability sampling**

Nonrandom Sampling

- **Every unit of the population does not have the same probability of being included in the sample.**
- **Open the selection bias**
- **Not appropriate data collection methods for most statistical methods**
- **Also known as nonprobability sampling**

Random Sampling Techniques

Simple Random Sample

Stratified Random Sample

- Proportionate
- Disportionate

Systematic Random Sample

Cluster (or Area) Sampling

Simple Random Sample

Number each frame unit from 1 to *N*.

Use a random number table or a random number generator to select *n* distinct numbers between 1 and *N*, inclusively.

Easier to perform for small populations

Cumbersome for large populations

Simple Random Sample: Sample Members

01 Alaska Airlines	11 DuPont	21 LTV
02 Alcoa	12 Exxon	22 Litton
03 Amoco	13 Farah	23 Mead
04 Atlantic Richfield	14 GTE	24 Mobil
05 Bank of America	15 General Electric	25 Occidental Petroleum
06 Bell Pennsylvania	16 General Mills	26 Penney
07 Chevron	17 General Dynamics	27 Philadelphia Electric
08 Chrysler	18 Grumman	28 Ryder
09 Citicorp	19 IBM	29 Sears
10 Disney	20 KMart	30 Time

$N = 30$

$n = 6$

Stratified Random Sample

Population is divided into nonoverlapping subpopulations called strata

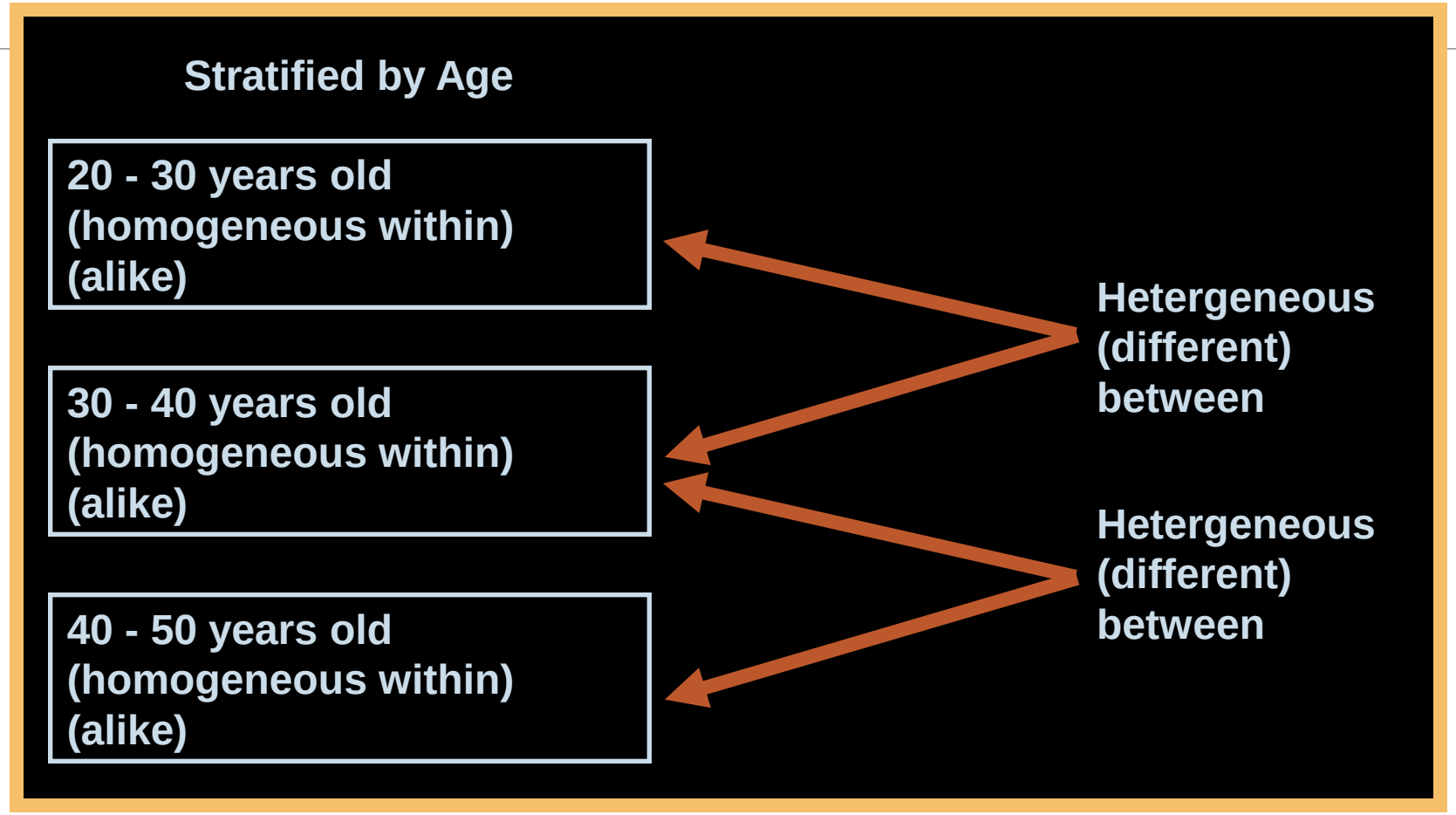
A random sample is selected from each stratum

Potential for reducing sampling error

Proportionate -- the percentage of thee sample taken from each stratum is proportionate to the percentage that each stratum is within the population

Disproportionate -- proportions of the strata within the sample are different than the proportions of the strata within the population

Stratified Random Sample: Population of FM Radio Listeners



Systematic Sampling

Convenient and relatively easy to administer

Population elements are an ordered sequence (at least, conceptually).

The first sample element is selected randomly from the first *k* population elements.

Thereafter, sample elements are selected at a constant interval, *k*, from the ordered sequence frame.

Systematic Sampling: Example

Consider that a hospital sees about 1 to 10,000 (*N* = 10,000) TB patients a year.

If we need a sample of fifty (*n* = 50) TB to study our phenomenom.

K(Interval)= 10,000/50 = 200

First sample element randomly selected from the first 200 purchase orders. Assume the 45th purchase order was selected.

Subsequent sample elements: 245, 445, 645, . . .

Cluster Sampling

Population is divided into nonoverlapping clusters or areas

Each cluster is a miniature, or microcosm, of the population.

A subset of the clusters is selected randomly for the sample.

If the number of elements in the subset of clusters is larger than the desired value of *n*, these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.

Cluster Sampling

Advantages

- More convenient for geographically dispersed populations
- Reduced travel costs to contact sample elements
- Simplified administration of the survey
- Unavailability of sampling frame prohibits using other random sampling methods

Disadvantages

- Statistically less efficient when the cluster elements are similar
- Costs and problems of statistical analysis are greater than for simple random sampling

Nonrandom Sampling

Convenience Sampling: sample elements are selected for the convenience of the researcher

Judgment Sampling: sample elements are selected by the judgment of the researcher

Quota Sampling: sample elements are selected until the quota controls are satisfied

Snowball Sampling: survey subjects are selected based on referral from other survey respondents

Errors

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.
- **Sampling Error** occurs when the sample is not representative of the population
- **Nonsampling Errors**
  - Missing Data, Recording, Data Entry, and Analysis Errors
  - Poorly conceived concepts , unclear definitions, and defective questionnaires
  - Response errors occur when people do not know, will not say, or overstate in their answers

Section 2: The importance of Sample size calculations
In a Clinical Research Setting

Sample size-Learning Objectives

✓Articulate the importance of sample size in clinical research

✓Know the key elements needed in calculating sample size

✓Appreciate how each of the key elements in calculating sample size influence the Size

✓Know some methods of adjustment of sample size to carter for dropouts in the study

Information Collection

1. Historical Data
   - **Pro:** Convenient; Save a lot of work
   - **Con:** Outdated; Different Objectives and Designs; Unknown Detailed Information
2. Census
   - **Pro:** reliable, accurate and comprehensive (e.g. Population census)
   - **Con:** Time consuming; requiring more resources; difficult to investigate all subjects in the population
3. Sampling
   - **Pro:** Efficient; Less risky; exploratory; informative
   - Caveats: Selection bias; misinterpretation; design flaw

Why Are we so interested in “How many” observations we need to prove our point?

Population

Sampling Plan

Conclusion

Inferential Statistics

Sample

Descriptive Statistics

Reasons for Appropriate sample Size

- Wrong conclusions
- Poor quality research (Errors)
  - Type II error can be minimized by increasing the sample size
- Waste of resources
- Loss of money
- Ethical problems
- Delay in completion

Excepts from Publications

- “Sample size calculation were based on a , which indicated that the baseline 5-year survival rate of D1 surgery was expected to be and an improvement in survival to (14% chance) with D2 resection would be a realistic expectation. Thus patients (200 in each arm) were to be randomized, providing to detect such a difference with p-value< . ”

Some Terminology

Significance level

Cut-off point for the p-value, below which the null hypothesis will be rejected and it will be concluded that there is evidence of an effect. Typically set at 5%.

One-sided and two-sided tests of significance

Two-sided tests should be used unless there is a very good reason for doing otherwise.

Power

Power is the probability that the null hypothesis will be correctly rejected i.e. rejected when there is indeed a real difference or association. It can also be thought of as "100 minus the percentage chance of missing a real effect" - therefore the higher the power, the lower the chance of missing a real effect. Power is typically set at 80% or 90% but not below 80%.

Effect size of clinical importance

This is the smallest difference between the group means or proportions (or odds ratio/relative risk closest to unity) which would be considered to be clinically or biologically important. The sample size should be set so that if such a difference exists, then it is very likely that a statistically significant result would be obtained.

Key Elements in Sample Size Calculation

- The level of statistical significance

- The anticipated clinical difference between treatment groups.

- The chance of detecting the anticipated clinical difference.

Steps in determining sample size

Determine the expected difference

↓

Find out the Standard deviations of both groups

↓

Set alpha error to be tolerated viz. P = 0.05

↓

Decide the power of the study desired viz. 80%, beta error 0.2

↓

Select the appropriate formula

↓

Calculate the sample size using the formula

↓

Give allowance for drop-out rate

↓

Give allowance for non-compliance of treatment if possible

Example of Formulae for sample size
Comparing two group means

(alpha=0.05, Beta=0.2, power 80%)

comparison

$$n = 16 \text{ X } (\text{S.D.}/M_1\text{-}M_2)^2$$

comparison

$$n = 8 \text{ X } (\text{S.D. of differences}/M_1\text{-}M_2)^2$$

Example of Formulae for sample size
Comparing two group proportions or %

(alpha=0.05, Beta=0.2, power 80%)

$$n = \frac{p1q1 + p2q2}{(p1-p2)^2} \times 8$$

P1= proportion of first group

P2= proportion of second group

Q1= 1-p1

Q2= 1-p2

Example
Scenario of Proportion-Cure rates

The cure rate of disease is 20% with a known drug treatment. It is claimed that yoga is better than the drug and a trial is to be conducted find out the truth. It is decided that a even ***10% increase in cure rate would be clinically important***. The alpha and beta were set at 0.05 and 0.2. The results will be analysed using Chi Square test. How many patients would be required for the trial?

Sample_Size _test of proportion

- Aim – To see whether yoga is better than standard drug x in curing the patient.
- Analysis type- **comparison of percentage** cure rate
- Parameters- cure rate **20%** vs **30%**
- No. of groups – 2
- p1=20 q1=80, p2=30 q2=70
- Set alpha=0.05, beta=0.2, Power=0.8
- Statistical formula to be used

$$n = \frac{p1q1 + p2q2}{(p1-p2)^2} \times 8$$

We need 296 patients

Cross sectional Study

- Cross-sectional studies include surveys
- People are studied at a “point” in time, without follow-up.
- Can combine a cross-sectional study with follow-up to create a cohort study.
- Can conduct repeated cross-sectional studies to measure change in a population.

Measure prevalence at “point” in time

- “Snapshot” of a population, a “still life”
- Can measure attitudes, beliefs, behaviors, personal or family history, genetic factors, existing or past health conditions, or anything else that does not require follow-up to assess.
- The source of most of what we know about the population

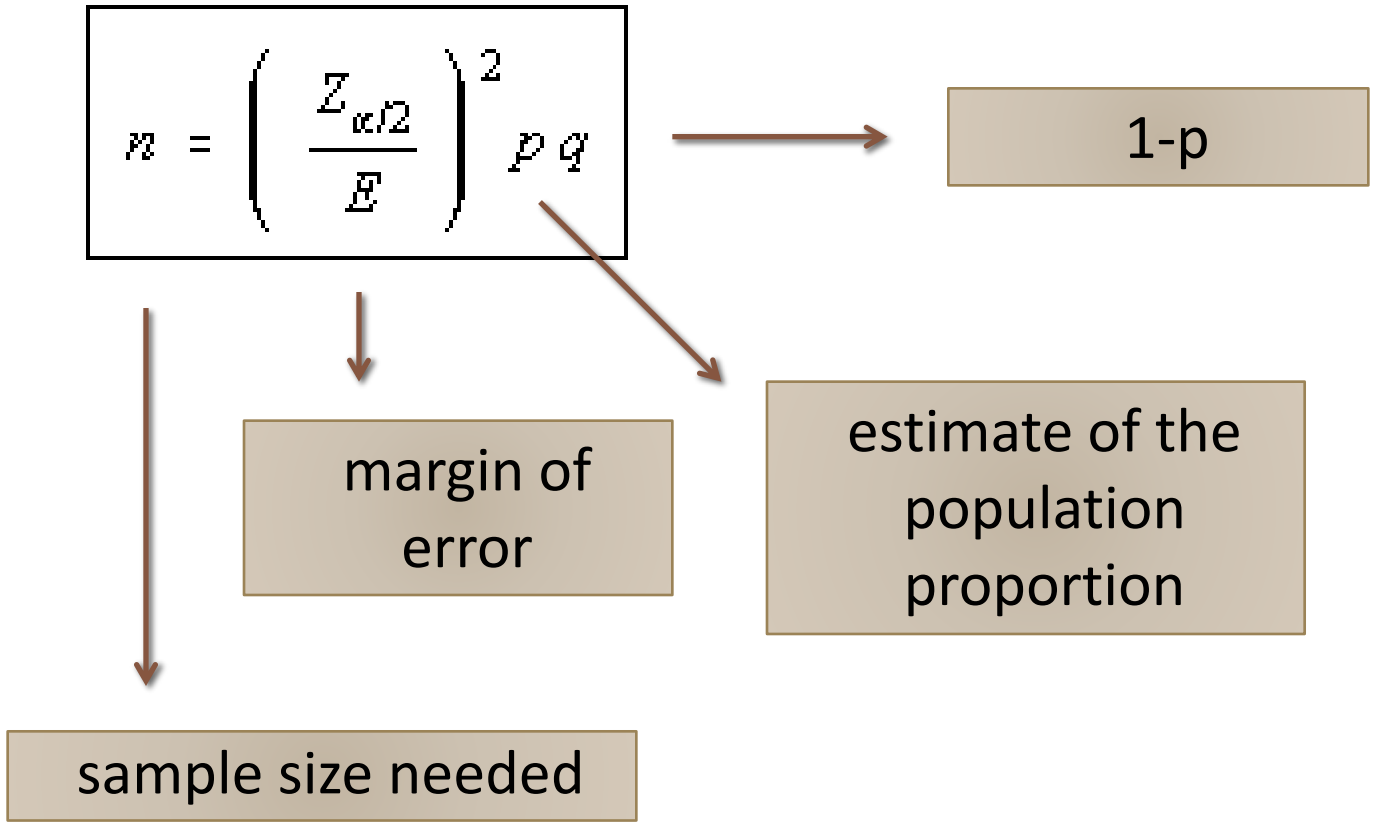
Sample Size

Formula for calculating sample size in a cross sectional study where proportion is the outcome measure

$$n = \frac{(1.96)^2 pq}{d^2}$$

n=size, p=proportion, q=1-P, d= effect size(how precise do you want to measure proportion)

Estimating a population proportion



Example

To estimate the proportion of hypertensive adults with a margin of error of 0.05 with 95% confidence. (p=20%)

Margin of error:0.05

Confidence: 95%

p = 20%

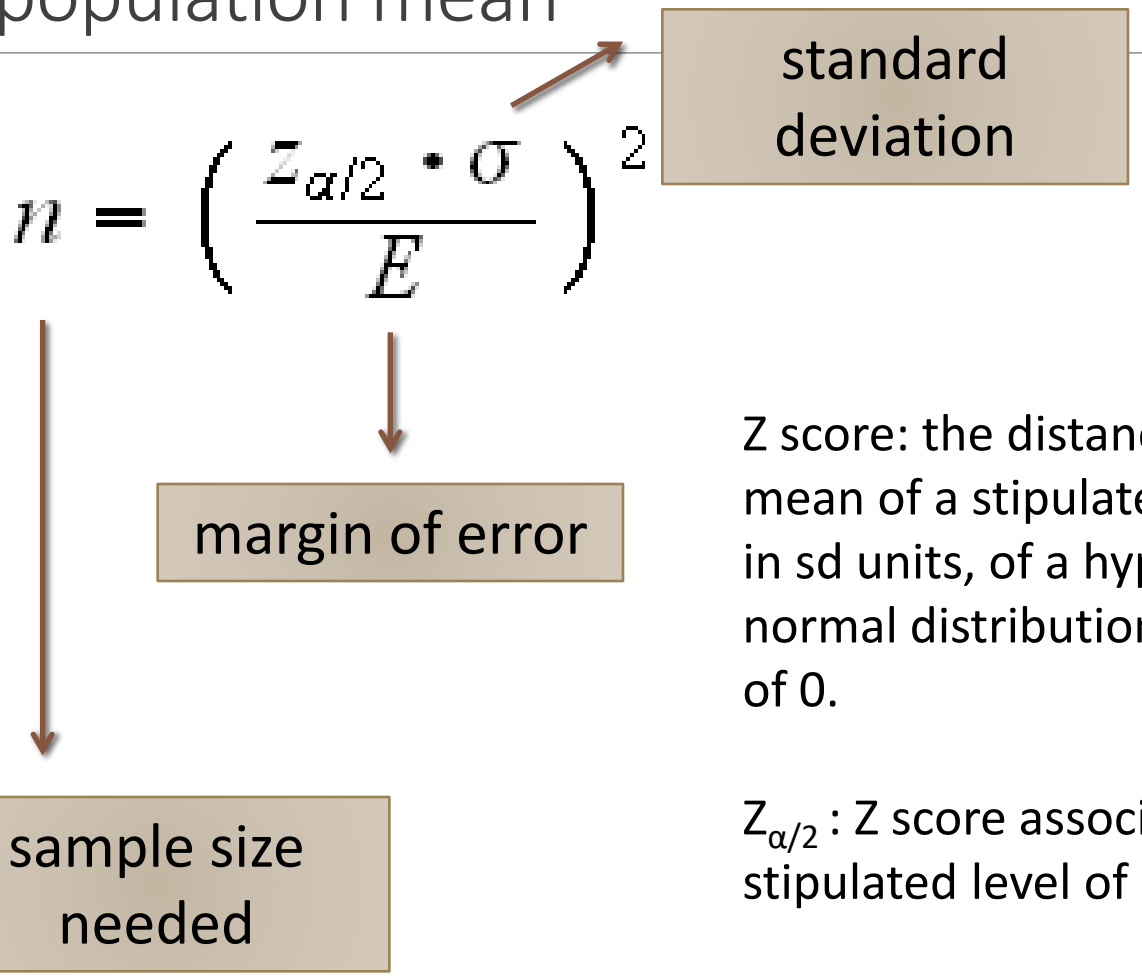
$$n = \left(\frac{Z_{\alpha/2}}{E} \right)^2 p\,q$$

$n = (1.96/0.05)^2\ (0.20 \times 0.80)$

n = 246

If we have no idea of p, then assume p=50%

Estimating a population mean



Z score: the distance from the mean of a stipulated probability, in sd units, of a hypothetical normal distribution with a mean of 0.

$Z_{\alpha/2}$: Z score associated with the stipulated level of α.

Example

To estimate the mean systolic blood for adults with a margin of error of 1 with 95% confidence. (sd=15mm-Hg)

Margin of error: 1

Confidence: 95%

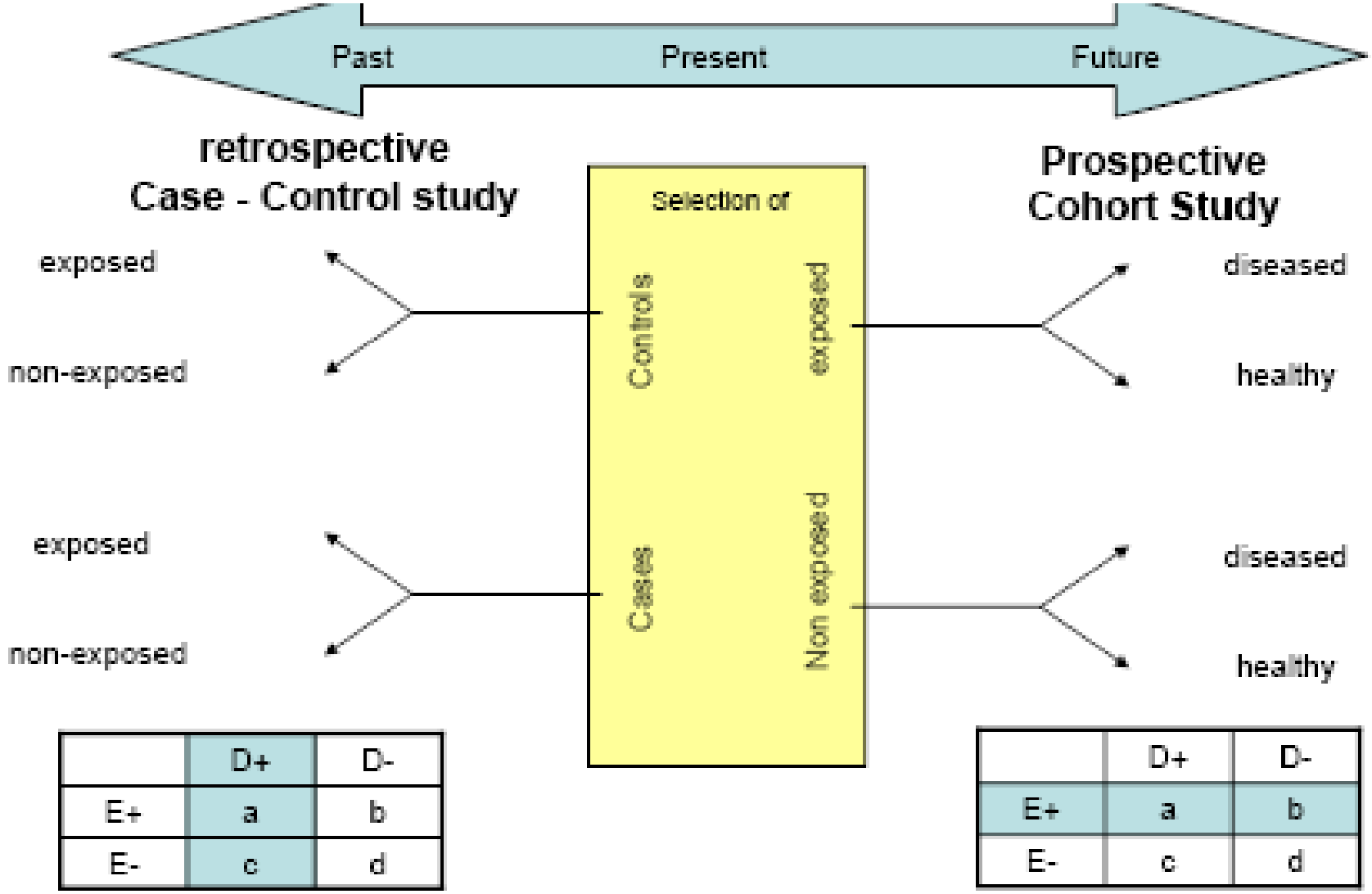
Sd: 15 mm-Hg

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

$n = (1.96 \times 15 / 1)^2$

n = 866

Case Control Studies/cohort studies



	D+	D-
E+	a	b
E-	c	d

	D+	D-
E+	a	b
E-	c	d

Bias in Case-Control studies

- **Selection bias**
  - Non-response
  - Detection bias
    - cases and controls are identified not independently of the exposure
- **Observation bias**
  - Recall Bias: Cases are more likely to remember exposure than controls

Bias-Cohort studies

- **Selection bias:**
  - Non-response during data collection
  - Losses to follow up
- **Misclassification on exposure or event**
  - Random
  - Systematic
- **Confounder**
  - Difference in other risk factors between exposed and non-exposed

Case Control Studies

- In case-control study the data are usually summarized by an odds ratio (OR), rather then difference between two proportions.
- If $p1$ and $p2$ are the proportions of cases and controls, respectively, exposed to a risk factor, then:

$$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

- If we know the proportion of exposure in the general population (p), the total sample size N for estimating an OR is:

$$N = \frac{(1+r)^2\left(Z_\alpha + Z_{1-\beta}\right)^2}{r(\ln OR)^2\, p(1-p)}$$

- Where $r = n_1 / n_2$ is the ratio of sample sizes for group 1 and group2; p is the prevalence of exposure in the controls; and OR is the hypothetical odds ratio. If $n1 = n_2$ (so that $r = 1$) then the fomula is reduced to:

$$N = \frac{4\left(Z_\alpha + Z_{1-\beta}\right)^2}{(\ln OR)^2\, p(1-p)}$$

Case Control Studies

- **Example**: The prevalence of vertebral fracture in a population is 25%. We are interested to estimate the effect of smoking on the fracture, with an odds ratio of 2, at the significance level of 5% (one-sided test) and power of 80%.
- The total sample size for the study can be estimated by:

$$N = \frac{4(1.64 + 0.85)^2}{(\ln 2)^2 \times 0.25 \times 0.75} = 275$$

More examples

Lets us look at more examples to get more comfortable with calculating sample sizes from Formulas

Q1) Sample size estimation for a descriptive study (single mean):

We want to estimate the mean systolic blood pressure of females in Yaounde.

The standard deviation is around 20 mmHg

We wish to estimate the true mean to within 10 mmHg with 95% confidence.

What is the required sample size ?

Solution 1 (using formula)

Sample size $n = Z^2_{(1-\alpha)}\, \sigma^2 / d^2$

- $s=20 \quad d=10 \quad (1-\alpha) = 0.95$
- $Z_{(1-\alpha)} = 1.96$ for 95% confidence level

$n = (1.96)^2 \times (20)^2 / (10)^2 = 15.37$

Since we cannot take 0.37 of a person, we round up to 16 women as our sample size.

descriptive study (single proportion):

We wish to estimate the proportion of Saudi males who smoke.

What sample size do we require to achieve a 95% confidence interval of width ± 5% (that is to be within 5% of the true value) ? In a study some years ago that found approximately 30% were smokers.

Solution 2 (using formula)

- Using the formula for sample size for a single proportion:
- Sample size $n = Z^2_{(1-\alpha)}\, p(1-p)/d^2$
  - $p=0.3$ $d=.05$
  - $(1-\alpha)=0.95$ $Z_{(1-\alpha)}= 1.96$ for 95% confidence level

- Then $n = (1.96)^2(0.3)(0.7)/(0.05)^2 = \mathbf{322.7}$

≈ 323

Q(3) Calculate Sample Size

- A new antihypertensive drug is to be tested against current treatment practice in people with systolic blood pressure > 160 mmHg and/or diastolic blood pressure >95 mmHg.
- It is felt that if the new drug can achieve blood pressure levels that are on the average 10 mmHg lower than those achieved using current treatment then it would be accepted by the medical community.
- The investigators would like at least 90% power and have chosen $\alpha = 0.01$ (two-sided) as the current therapy is quite acceptable and they want to be sure that the new therapy is superior before switching over. Blood pressure measurements have a standard deviation of 20 mmHg.

Solution 3 (using formula)

Sample size $n = \underline{2\,\sigma^2\,(Z_{(1-\alpha)} + Z_{(1-\beta)})^2}/d^2$

Where:

- d=10 σ=20 α=0.01 $Z_{(1-\alpha)}$ = 2.58 β=*(1-power)= 1-0.9=* **0.10**
  $Z_{(1-\beta)}$ = 1.28

Plug into formula:

$n = 2 \times 20^2\,(2.58+1.28)^2\,/(10)^2 = 119.2 \approx$ ***120***

Question 4

A standard regimen has an efficacy of 80% and a new regimen has been claimed to be 90% effective.

What is the sample size required to test whether the new treatment is really effective at 5% level with 90% power ?

Solution 4 (using formula)

Sample size calculation for proportion:

$$n = \frac{(Z_{(1-\alpha)}+Z_{(1-\beta)})^2 [p_1(1-p_1)+ p_2(1-p_2)]}{(p_1-p_2)^2}$$

- p1=80% p2=90% α=0.05 (1-α)=0.95
- $Z_{(1-\alpha)}$= 1.96 for 95% confidence level
- (Power=1-β)=0.90 $Z_{(1-\beta)}$= 1.282 for 90% power

n= $(1.96+1.282)^2((0.8x0.2)+(0.9x0.1))$ **/** $(0.8-0.9)^2$= **263** patients for each treatment

Total sample size: 263x2= *526*

Sample size calculation in Cohort studies

- Same procedure and formula like in the case control setting

- The difference is on the fact that instead of using the Odds Ratio(OR) , it uses the Relative Risk(RR)

Sample size, Power,level of significance and Effect size

- Effect size$\uparrow \rightarrow$Size $\downarrow$
- The power 1- $\beta\uparrow \rightarrow$Size $\uparrow$
- The level of $\alpha\downarrow \rightarrow$Size $\uparrow$

Think about Numbers at the Begining

- Sample size consideration should be thought of as you write your research proposal

- Requirement of Ethics committees and Journals

- It is the bases of whether you would be able to prove what you want to prove

- It could be influenced by resources

- Always take care to consider your design and look out for drop out during the course of the study and make adjustment for this in your sample size calculations

What would you need to measure your outcome???

- Case Report form.
  - Paper based
  - Electronic
- Questionnaires
  - Interviews
  - Mail
  - phone

Data collection

- Hospital records
- Observations
- Lab records
- Interviewing
- Literature reviews

Data Processing

- Electronic copy of data
- Clean data
- Determine whether data answers your question
- Interprete results

Variables

- A variable is an object, characteristic or property that can have different values in different places, persons, or things.
- A quantitative variable
  - Examples: Heart rate, heights, weight, age, size of tumor, volume of a dose.
- A qualitative (categorical) variable is characterized by its inability to be measured but it can be sorted into categories. •Examples: gender, race, drug name, disease status.

We will dig in next week

THANK YOU AND SEE YOU on Tuesday